Reference 3

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 11-039313
(43)Date of publication of application : 12.02.1999

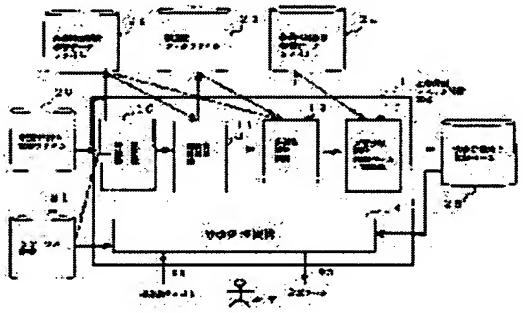
(51)Int.Cl. G06F 17/30

(21)Application number : 09-198113
(22)Date of filing : 24.07.1997

(71)Applicant : NIPPON TELEGR & TELEPH CORP <NTT>
(72)Inventor : YAMAZAKI TAKEFUMI

(54) AUTOMATIC DOCUMENT CLASSIFICATION SYSTEM, DOCUMENT CLASSIFICATION ORIENTED KNOWLEDGE BASE CREATING METHOD AND RECORD MEDIUM RECORDING ITS PROGRAM

(57)Abstract:
PROBLEM TO BE SOLVED: To dissolve the polysemy of a semantic category and to improve precision more by selecting and utilizing only an optimum semantic category that a word has at the time of generating learning data.
SOLUTION: A characteristic extraction mechanism 10 analyzes a classification tagged document set of a file 20, examines the semantic category of a word and outputs an extracted characteristic vector and data of a classification category to which its text belongs to a learning data file 22 before polysemy dissolution. A polysemy dissolution mechanism 12 is inputted from the file 22 and an association degree data file 23 and outputs a polysemy dissolution result to a learning data file 24 after polysemy dissolution. A document classification oriented knowledge base creation mechanism 13 is inputted from the file 24 and outputs the weight between a characteristic and a classification to a document classification oriented knowledge base 25. A classification processing mechanism 14 is inputted from a thesaurus dictionary 21 and the base 25 and outputs a classification tag that corresponds to a new text inputted by a user.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

Reference 3

(19)日本国特許庁（ＪＰ）　(12)**公開特許公報**（Ａ）　(11)特許出願公開番号

特開平11－39313

(43)公開日　平成11年(1999)２月12日

(51)Int.Cl.[6]	識別記号	ＦＩ	
Ｇ０６Ｆ　17/30		Ｇ０６Ｆ　15/401	３１０Ｄ

審査請求　未請求　請求項の数３　ＯＬ　（全　８　頁）

(21)出願番号　特願平9－198113

(22)出願日　平成９年(1997)７月24日

(71)出願人　000004226
日本電信電話株式会社
東京都新宿区西新宿三丁目19番２号

(72)発明者　山崎　毅文
東京都新宿区西新宿三丁目19番２号　日本電信電話株式会社内

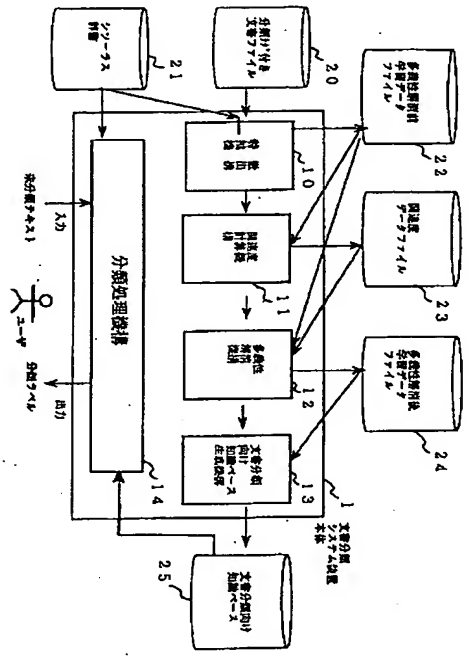
(74)代理人　弁理士　鈴木　誠

(54)【発明の名称】　文書自動分類システム、文書分類向け知識ベース生成方法及びそのプログラムを記録した記録媒体

(57)【要約】

【課題】　意味カテゴリーの多義性を解消し、高精度の文書自動分類システムを提供する。

【解決手段】　分類タグ付き文書集合を入力し、シソーラス辞書を用いて、該文書を特徴付ける単語とその意味カテゴリーと該文書の分類カテゴリーからなる学習データを生成する手段、該学習データについて、分類カテゴリーと特徴間の関連度を計算する手段、該関連度に基づいて、前記学習データから不適切な特徴を除去して新たな学習データを生成する手段、該新たな学習データについて、分類カテゴリーと特徴間の重みを計算して文書分類向け知識ベースを生成する手段を設ける。



【特許請求の範囲】
【請求項１】　シソーラス辞書を用いた文書自動分類システムにおいて、分類タグ付き文書集合とシソーラス辞書を入力とし、文書を特徴付ける単語とその意味カテゴリーと当該文書の分類カテゴリーから構成される学習データを生成する手段と、前記生成された学習データを入力とし、分類カテゴリーと特徴間の関連度を計算する手段と、前記計算された関連度と前記学習データを入力とし、学習データから不適切な特徴を除去して新たな学習データを生成する手段と、前記生成された新たな学習データを入力とし、分類カテゴリーと特徴間の重みを計算して文書分類向け知識ベースを生成する手段と、未分類の文書を入力とし、前記知識ベースを元に対応する分類カテゴリーを出力する手段とを有することを特徴とする文書自動分類システム。
【請求項２】　シソーラス辞書を用いた文書自動分類システムにおける文書分類向け知識ベースを生成する方法であって、
分類タグ付き文書集合とシソーラス辞書を入力して、文書を特徴付ける単語とその意味カテゴリー及び当該文書の分類カテゴリーから構成される学習データを生成し、
前記学習データについて、分類カテゴリーと特徴間の関連度を計算し、
前記関連度に基づき、前記学習データから不適切な特徴を除去して新たな学習データを生成し、
前記新たな学習データにより、分類カテゴリーと特徴間の重みを計算して、文書分類向け知識ベースを生成することを特徴とする文書分類向け知識ベース生成方法。
【請求項３】　シソーラス辞書を用いた文書自動分類システムにおける文書分類知識ベースを生成する処理のプログラムを記録したコンピュータ読み取り可能な記録媒体であって、
分類タグ付き文書集合を読み取る処理と、
前記分類タグ付き文書集合を解析して、当該文書を特徴付ける単語を抽出し、シソーラス辞書を参照して該単語の意味カテゴリーを調べ、前記単語と意味カテゴリーからなる特徴ベクトル及び当該文書の分類カテゴリーから構成される学習データを生成する処理と、
前記学習データについて、分類カテゴリーと特徴間の関連度を計算する処理と、
前記関連度に基づき、前記学習データから不適切な特徴を除去して新たな学習データを生成する処理と、
前記新たな学習データにより、分類カテゴリーと特徴間の重みを計算して、文書分類向け知識ベースを生成する処理と、を含むプログラムを記録したコンピュータ読み取り可能な記録媒体。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は自然言語処理の技術分野に係わり、詳しくは、文書自動分類システム、分類タグ付き文書集合から文書分類向け知識ベースを生成する方法及びそのプログラムを記録した記録媒体に関するものである。
【０００２】
【従来の技術】文書分類とは、テキストをその意味内容に基づいて、予め決められた分類カテゴリー（例えば、政治、スポーツ、経済等）に分類することであり、大量の電子化されたテキストが流通するようになった現在、テキストの効率的検索／利用において、重要な課題となっている。この文書分類作業の自動化を実現するのが文書自動分類システムである。このシステムは、予め、カテゴリー分類されたテキスト（ここでは、これを「分類タグ付き文書集合」と呼ぶ）を利用した適当な分類器（文書分類向け知識ベース）を作成し、その分類器に基づいて新たに入力されたテキストのカテゴリー分類を行う。
【０００３】従来の文書自動分類システムは、分類タグ付き文書集合から、分類のキーとなる単語を抜き出し、それらの単語と予め付与された分類カテゴリーとの結び付きの強さである関連度を学習して、分類器を生成する。同じ分類カテゴリーを持つテキストには、その分野を特徴づける単語が決まって頻出するので、その単語とその指定した分類カテゴリーとの関連度は大きくなり、そうでない単語と分類カテゴリーとの関連度は小さくなる。分類器は、単語、分類カテゴリー、及びこの単語と分類カテゴリーとの関連度を表わすリンクから構成される。
【０００４】分類器の例を図２に示す。図２で示す通り、特徴ノードはテキストで頻出する単語、分類ノードは文書集合が持つ分類カテゴリーから構成され、特徴ノードと分類ノード間のリンク上の数値が、関連度を表わす。関連度の学習方法として、カイ２乗計算手法等、様々な手法が考えられる。
【０００５】新たなテキストが入力されると、まず、そのテキスト上に現われる単語を抽出し、次に分類器上の単語と分類タグ間の関連度から、そのテキストの各分類カテゴリーへの関連度を計算する。この値が、ある決められた閾値を超えるもの或いは最も大きい値を持つものを、そのテキストの分類カテゴリーと判断する。
【０００６】
【発明が解決しようとする課題】上記従来技術によって、分類タグ付き文書集合から自動的に分類器が生成できるが、特徴量ノードを構成するものが、分類タグ付き文書集合上に出現する単語に限られるので、新たに入力される分類対象テキストが、特徴ノードにない未知単語を含む場合、そのテキストの分類精度が悪くなってしまう。この問題を解決する一つの方法として、特徴ノードの生成において、単語のみでなく、シソーラス利用によって得られる単語の意味カテゴリーも合わせて、利用する手法が考えられる。意味カテゴリーは、単語よりも一

般的な概念で、通常複数の単語が共通の意味カテゴリーを持つので、この意味カテゴリーによる特徴ノードの生成で、未知単語の問題が解決できる。

【０００７】しかし、一方で、シソーラス利用によって得られる各単語の意味カテゴリーの数は通常一つとは限らない。例えば、単語「路線」は、政治の分野では「方針」という意味で使われ、また、交通の分野では「道筋」という意味で使われる。このような多義語は、シソーラス上で複数の意味カテゴリーを持つ。分類器を生成する学習段階で、シソーラスを利用して特徴ノードを生成する場合に、多義語から得られる複数の意味カテゴリーを、文脈に応じて絞らずに全て利用すると、特徴ノードと分類ノードとの関連度が誤って計算されてしまい、分類精度が悪くなる可能性がある。特徴ノードにシソーラス上の意味カテゴリーを利用する場合は、意味カテゴリーを文脈に応じて正確に特定する、所謂多義解消が必要である。

【０００８】本発明の目的は、意味カテゴリーの多義性を解消し、より精度の高い文書自動分類システム、その文書分類向け知識ベースの生成方法及びそのプログラムを記録した記録媒体を提供することにある。

【０００９】

【課題を解決するための手段】本発明の文書自動分類システムは、分類タグ付き文書集合とシソーラス辞書を入力し、文書を特徴付ける単語とその意味カテゴリーと当該文書の分類カテゴリーから構成される学習データを生成する手段と、前記生成された学習データから分類カテゴリーと特徴間の関連度を計算する手段と、前記計算された関連度と前記学習データを入力し、学習データから不適正な意味カテゴリーを除去して新たな学習データを生成する手段と、前記生成された新たな学習データを入力とし、分類カテゴリーと特徴間の関連度を計算して文書分類向け知識ベースを生成する手段と、未分類の文書を入力とし、前記知識ベースを元に対応する分類カテゴリーを出力する手段とを有することを特徴とする。

【００１０】本発明の文書分類向け知識ベース生成方法は、分類タグ付き文書集合とシソーラス辞書を入力して、文書を特徴付ける単語とその意味カテゴリー及び当該文書の分類カテゴリーから構成される学習データを生成し、前記学習データについて、分類カテゴリーと特徴間の関連度を計算し、前記関連度に基づき、前記学習データから不適切な特徴を除去して新たな学習データを生成し、前記新たな学習データにより、分類カテゴリーと特徴間の重みを計算して、文書分類向け知識ベースを生成することを特徴とする。

【００１１】本発明のコンピュータ読み取り可能な記録媒体は、分類タグ付き文書集合を読み取る処理と、前記分類タグ付き文書集合を解析して、当該文書を特徴付ける単語を抽出し、シソーラス辞書を参照して該単語の意味カテゴリーを調べ、前記単語と意味カテゴリーからなる特徴ベクトル及び当該文書の分類カテゴリーから構成される学習データを生成する処理と、前記学習データについて、分類カテゴリーと特徴間の関連度を計算する処理と、前記関連度に基づき、前記学習データから不適切な特徴を除去して新たな学習データを生成する処理と、前記新たな学習データにより、分類カテゴリーと特徴間の重みを計算して、文書分類向け知識ベースを生成する処理とを含むことを特徴とする。

【００１２】意味カテゴリーの多義性解消は、「意味カテゴリーは、その対象テキストが属する分類カテゴリーに関連のある他の単語からも生成される」という性質に基づいて、テキストの分類カテゴリー情報と事前に得られる特徴と該当分類カテゴリー間の関連度の利用により行える。例えば、多義語「路線」は、あるシソーラス上で、「Ｓ交通路」「Ｓ形勢」の２つの意味カテゴリーを持つ。ここでは、意味カテゴリーを単語と区別するため、意味カテゴリーの先頭に文字Ｓを付与する。「Ｓ形勢」は、分類カテゴリーが「政治」で頻出する単語「動向」「非常事態」からも生成される意味カテゴリーであり、また「Ｓ交通路」は、分類カテゴリーが「交通」で頻出する単語「新幹線」「東海道」からも生成される。よって、「政治」「交通」の分類カテゴリーを持つカテゴリーを持つテキストから、「Ｓ交通路」は「交通」と強い関連があり、「Ｓ形勢」は「政治」と強い関連を持つことが解るはずである。この関連度を利用すれば、対象テキスト中に単語「路線」が現われた時、そのテキストの分類カテゴリーが「政治」であれば、意味カテゴリーとして「Ｓ形勢」を選択し、「交通」であれば、「Ｓ交通路」を選択して用いることが可能である。

【００１３】このように、シソーラス利用の文書自動分類システムにおいて、最適な意味カテゴリーを選択する、多義を解消する処理プログラムを組み込むことにより、学習データ中のノイズが減少し、より精度の高い文書分類向け知識ベースを生成することが可能になる。

【００１４】

【発明の実施の形態】以下、図面を用いて、本発明の一実施例を説明する。図１は、本発明の一実施例に係わる文書自動分類システムのブロック図である。文書自動分類システムは、特徴抽出機構１０、関連度計算機構１１、多義性解消機構１２、文書分類向け知識ベース生成機構１３、及び未分類の文書に分類ラベルを出力する分類処理機構１４から装置本体１と、分類タグ付き文書ファイル２０、シソーラス辞書２１、多義性解消前の学習データファイル２２、関連度データファイル２３、多義性解消後の学習データファイル２４、文書分類向け知識ベース２５等を格納する外部記憶装置群で構成される。装置本体１は、ＣＰＵ、ＲＡＭ、内蔵ハードディスクなどで構成される所謂コンピュータである。

【００１５】図１の構成において、特徴抽出機構１０は、分類タグ付き文書ファイル２０及びシソーラス辞書

２１を入力として、文書を表わス特徴及びその分類カテゴリー情報を多義性解消前の学習データファイル２２に出力する。関連度計算機構１１は、多義性解消前の学習データファイル２２を入力として、特徴と分類カテゴリー間の関連度情報を関連度データファイル２３に出力する。多義性解消機構１２は、多義性解消前の学習データファイル２２及び関連度データファイル２３を入力として、多義性解消結果を多義性解消後の学習データファイル２４に出力する。文書分類向け知識ベース生成機構１３は、多義性解消後の学習データファイル２４を入力として、再計算後の特徴と分類カテゴリー間の関連度を文書分類向け知識ベース２５に出力する。分類処理機構１４は、文書分類向け知識ベース２５及びユーザが入入力した新たなテキストを入力として、入力されたテキストに対応する分類カテゴリーを出力する。以下に、各機構１０、１１、１２、１３、１４の構成および動作を詳述する。

【００１６】〈特徴抽出機構１０〉特徴抽出機構１０は、ファイル２０の入力として与えられた分類タグ付き文書集合を解析し、品詞が名詞、固定名詞である単語を抜き出し、その後、シソーラス辞書２１を参照して、それらの単語の意味カテゴリーを調べる。そして、一つのテキストから、そのテキスト中に含まれる単語と意味カテゴリーからなる特徴ベクトルを抽出し、該特徴ベクトルとそのテキストの属する分類カテゴリーからなるデータを多義性解消前の学習データファイル２２に出力する。

【００１７】図３に、特徴抽出機構１０の一実施例の構成図を示す。本特徴抽出機構１０は、形態素解析部１０１、単語抽出部１０２、意味カテゴリー付与部１０３からなる。

【００１８】形態素解析部１０１は、ファイル２０から入力されたテキストについて形態素解析を行い、単語抽出とその品詞付けを行う。単語抽出部１０２は、必要な品詞（名詞、固有名詞）の単語を選び出す。意味カテゴリー付与部１０３は、単語抽出部１０２で選ばれた単語について、利用するシソーラス辞書２１を調べて、その意味カテゴリーを付与し、単語と意味カテゴリーからなる特徴ベクトルと該テキストの属する分類カテゴリーからなるデータ（多義性解消前の学習データ）をファイル２２に出力する。

【００１９】図４は、本特徴抽出機構１０での具体的処理例を示したものである。ここで、意味カテゴリーの先頭には文字「Ｓ」を付与し、単語と区別する。例えば、単語「路線」に対し、シソーラス辞書２１において「交通路」「形勢」の二つの意味カテゴリーを持つ場合、「Ｓ交通路」、「Ｓ形勢」を付加する。「Ｓ党」、「Ｓ団体」も同様である。

【００２０】〈関連度計算機構１１〉関連度計算機構１１は、多義解消前の学習データファイル２２を入力として、特徴と分類カテゴリー間の関連度情報を関連度データファイル２３に出力する。

【００２１】図５に、関連度計算機構１１の一実施例の構成図を示す。本関連度計算機構１１は、事例数カウンターモジュール１１１、カイ２乗計算モジュール１１２、関連度格納モジュール１１３から構成される。

【００２２】事例数カウンターモジュール１１１は、多義解消前の学習データファイル２２を入力として、各分類カテゴリー毎に、各特徴に対して、次の４つの値をカウントする。即ち、該当特徴が出現するテキスト集合中で、該当分類カテゴリーを持つテキスト数「Ｎｒ＋」、持たないテキストの数「Ｎｎ＋」、また、該当特徴が出現しないテキスト集合中で、該当分類カテゴリーを持つテキスト数「Ｎｒ－」、持たないテキストの数「Ｎｎ－」をそれぞれ求める。

【００２３】カイ２乗計算モジュール１１２は、上記４つの値を用いて、次の式に基づいて、カイ２乗値（関連度）を計算する。但し、Ｎは全事例数を表す。

【００２４】

【数１】

$$\chi^2=\frac{N\,(N_{r+}N_{n-}-N_{r-}N_{n+})^2}{(N_{r+}+N_{r-})\,(N_{n+}+N_{n-})\,(N_{r+}+N_{n+})\,(N_{r-}+N_{n-})}$$

【００２５】関連度格納モジュール１１３は、カイ２乗計算モジュール１１２の計算結果（各分類カテゴリー毎、各特徴毎のカイ２乗値）を関連度データファイル２３に格納する。

【００２６】〈多義性解消機構１２〉多義性解消機構１２は、多義性解消前の学習データファイル２２及び関連度データファイル２３を入力として、多義性解消結果を多義性解消後の学習データファイル２４に出力する。

【００２７】図６は、多義性解消機構１２の一実施例の構成図を示す。本多義性解消機構１２は、多義語選択モジュール１２１、関連度参照モジュール１２２、最適意味カテゴリー選択モジュール１２３から構成される。

【００２８】多義語選択モジュール１２１は、多義性解消前の学習データファイル２２を入力として、意味カテゴリーが複数付与されている多義語を探しだし、その意味カテゴリー及びそのテキストが属する分類カテゴリーを出力する。関連度参照モジュール１２２は、関連度データファイル２３を入力として、モジュール１２１で得られた意味カテゴリーと分類カテゴリーから、それらに対応する関連度を出力する。最適意味カテゴリー選択モジュール１２３は、モジュール１２２で得られた関連先を元に、関連度の最も大きい値をもっものを最適意味カテゴリーとして選択し、多義性解消後の学習データファイル２４に出力する。

【００２９】例えば、図４の場合、多義語「路線」に対応する「Ｓ交通路」と「Ｓ形勢」の２つの意味カテゴリーのうち、「Ｓ形勢」が最適意味カテゴリーとして選択される。

【００３０】〈文書分類向け知識ベース生成機構１３〉文書分類向け知識ベース生成機構１３は、多義性解消後の学習データファイル２４を入力として、特徴と分類タグ間の重みを文書分類向け知識ベース２５に出力する。実施例の一例として、ここでは、文書分類を線形分類モデルに基づいて行ない、線形モデルにおける重みの学習に誤り駆動型学習アルゴリズムを用いるとする。線形分類モデルは、特徴集合ノードと分類カテゴリーノードとから構成されており、入力された事例がその分類カテゴリーに属するか否かの判定を、入力事例のもつ特徴ノードと分類カテゴリー間の重みの合計が、決められた閾値を超えるか否かによって行う。線形分類モデルの例を図７に示す。なお、線形分類モデル、誤り駆動型学習アルゴリズムについては、例えば「N. Littlestone, "Learning quickly when irrelevant attributes abound : A new linearthreshold algorithm", Machine Learning, No. 2, pp２８５－３１８, １９８８. 」に記述されている。

【００３１】図８は、誤り駆動型学習アルゴリズムを用いた、文書分類向け知識ベース生成機構１３の一実施例の構成図を示す。本文書分類向け知識ベース生成機構１３は重み初期値設定モジュール１３１、重み合算モジュール１３２、正解判定モジュール１３３、重み更新モジュール１３４から構成される。

【００３２】重み初期値設定モジュール１３１は、多義性解消後の学習データファイル２４を入力として、学習データに出現する特徴と分類カテゴリー間をある値に初期化し、文書分類向け知識ベース２５に出力する。重み合算モジュール１３２は、多義性解消後の学習データファイル２４、文書分類向け知識ベース２５を入力として、各入力事例毎に、事例に出現する特徴から、各分類カテゴリーに対するスコアを重みの合計として計算する。正解判定モジュール１３３は、前モジュールで計算されたスコアが、決められた閾値を超えるか否かを判定し、分類カテゴリーを割り当てる。重み更新モジュール１３４は、前モジュールが割り当てた分類カテゴリーが正解の分類カテゴリーと異なる場合のみ、特徴と分類カテゴリー間の重みを更新し、文書分類向け知識ベース２５に出力する。

【００３３】〈分類処理機構１４〉分類処理機構１４は、シソーラス辞書２１、文書分類向け知識ベース２５及びユーザが入力した新たなテキストを入力として、入力されたテキストに対応する分類タグを出力する。

【００３４】図９は、分類処理機構１４の一実施例の構成図を示す。本分類処理機構１４は特徴抽出モジュール１４１、重み合算モジュール１４２、分類カテゴリー生成モジュール１４３から構成される。

【００３５】特徴抽出モジュール１４１は、ユーザが入力した新たな未分類テキストを形態素解析し、名詞、固定名詞である単語を選択し、それらの持つ意味カテゴリーを、シソーラス辞書２１を参照して付与し、特徴ベクトルを生成する。本モジュール１４１は、前記特徴抽出機構１１と基本的に同じものである。重み合算モジュール１４２は、モジュール１４１で生成された特徴ベクトルと文書分類向け知識ベース２５を入力として、各分類カテゴリーに対するスコア計算を行う。分類カテゴリー生成モジュール１４３は、前モジュールで計算されたスコアがある決められた閾値以上である分類カテゴリーを出力する。

【００３６】以上、本発明の一実施例に係わる文書自動分類システムについて説明したが、図１において、文書分類向け知識ベースの生成に関係する特徴抽出機構１０、関連度計算機構１１、多義性解消機構１２及び文書分類向け知識ベース生成機構１３の処理プログラムは一つにまとめてもよい。図１０は、その処理フローを示したもので、処理１０１０～１０１３は図１の各機構１０～１３に対応する。処理１０１０～１０１３のプログラムは、あらかじめＣＤ－ＲＯＭ等の記録媒体に記録しておき、該プログラムをコンピュータにロードすることにより、先に説明した図１の各機構１０～１３と同様の処理が実現する。

【００３７】

【発明の効果】以上説明したように、本発明によれば、分類タグ付き文書集合とシソーラス辞書から、分類器（文書分類向け知識ベース）の生成に利用する学習データを作成する際に、単語の持つ意味カテゴリーを最適な意味カテゴリーのみを選んで利用するので、従来手法による多義性解消をしない場合に比べて、より精度の高い分類器を生成することができる。

【図面の簡単な説明】

【図１】本発明の一実施例の文書自動分類システムの全体構成図である。

【図２】分類器の一例を示す図である。

【図３】図１の特徴抽出機構の一実施例の構成図である。

【図４】特徴抽出機構の具体的処理例を示す図である。

【図５】図１の関連度計算機構の一実施例の構成図である。

【図６】図１の多義性解消機構の一実施例の構成図である。

【図７】線形分類モデルの一例を示す図である。

【図８】図１の文書分類向け知識ベース生成機構の一実施例の構成図である。

【図９】図１の分類処理機構の一実施例の構成図である。

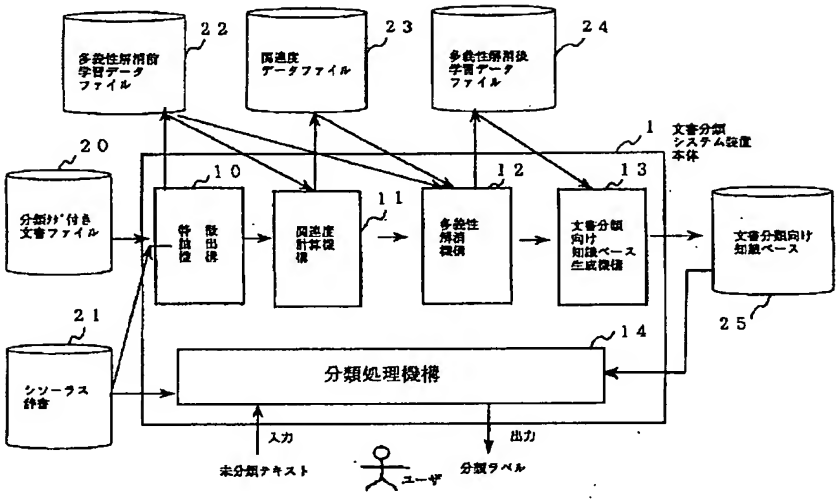
【図１０】本発明の文書分類向け知識ベース生成方法の

概略処理フロー図である。

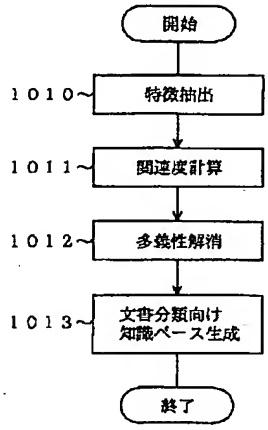
【符号の説明】

１　文書自動分類装置本体
２　外部記憶装置群
１０　特徴抽出機構
１１　関連度計算機構
１２　多義性解消機構
１３　文書分類向け知識ベース生成機構
１４　分類処理機構
２０　分類タグ付き文書ファイル
２１　シソーラス辞書
２２　多義性解消前の学習データファイル
２３　関連度データファイル
２４　多義性解消後の学習データファイル
２５　文書分類向け知識ベース

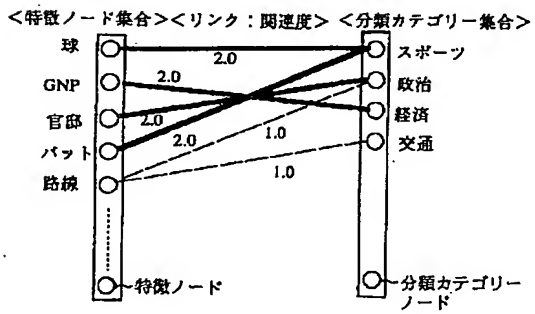
【図１】



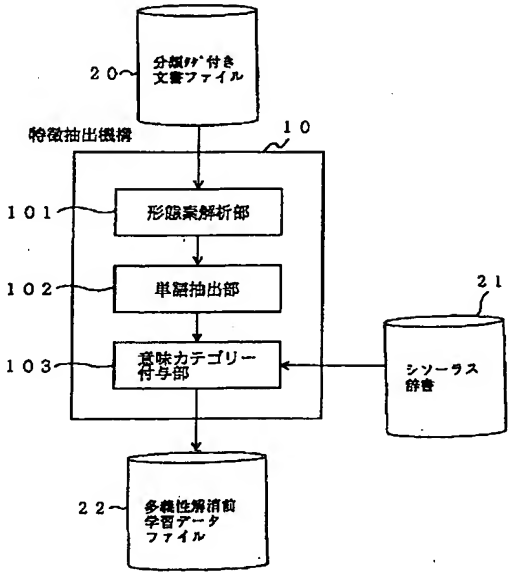
【図１０】



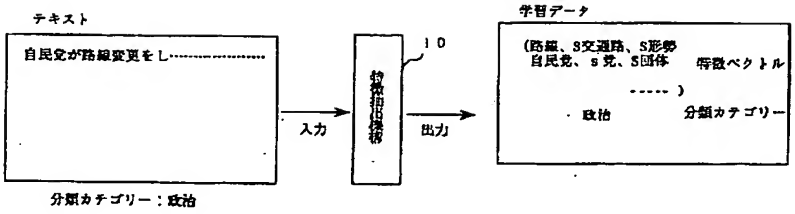
【図２】



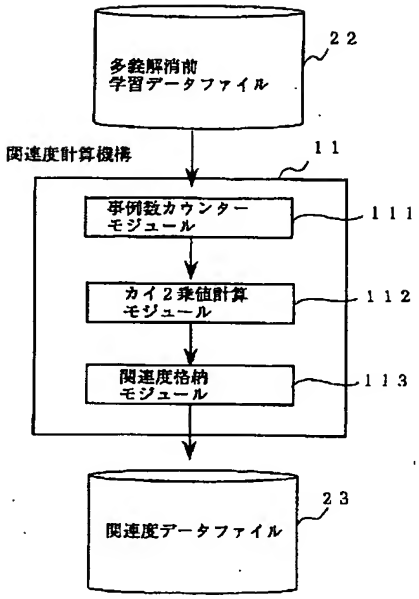
【図３】



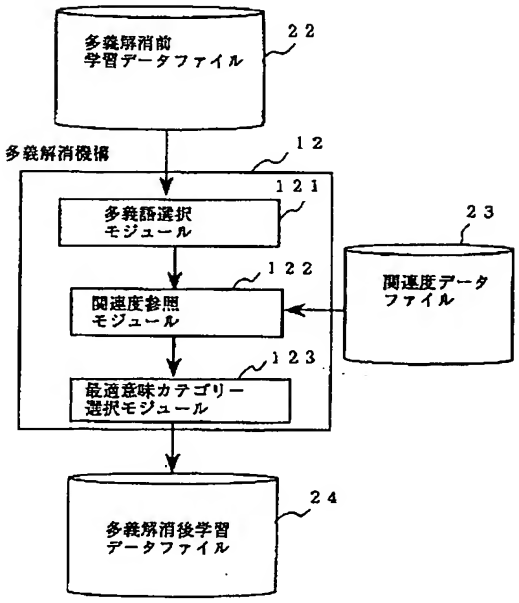
【図４】


テキスト
自民党が路線変更をし……
分類カテゴリー：政治
入力
10
出力
学習データ
（路線、S交通路、S形勢
自民党、S党、S団体
特徴ベクトル
……）
政治
分類カテゴリー

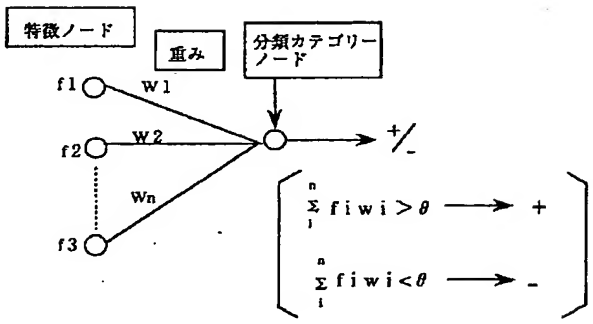
【図５】


多義解消前
学習データファイル
22
関連度計算機構
11
事例数カウンター
モジュール
111
カイ２乗値計算
モジュール
112
関連度格納
モジュール
113
関連度データファイル
23

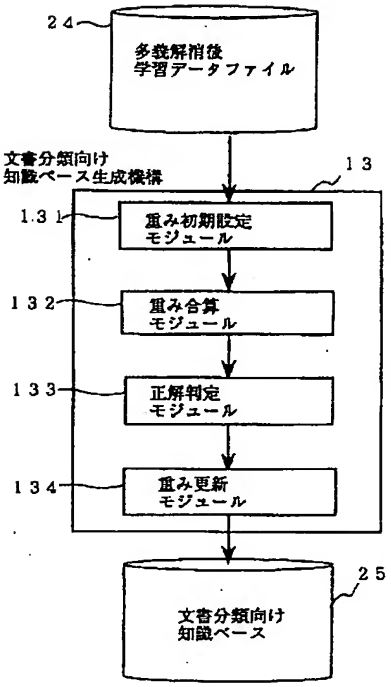
【図６】


多義解消前
学習データファイル
22
多義解消機構
12
121
多義語選択
モジュール
122
関連度参照
モジュール
123
最適意味カテゴリー
選択モジュール
23
関連度データ
ファイル
多義解消後学習
データファイル
24

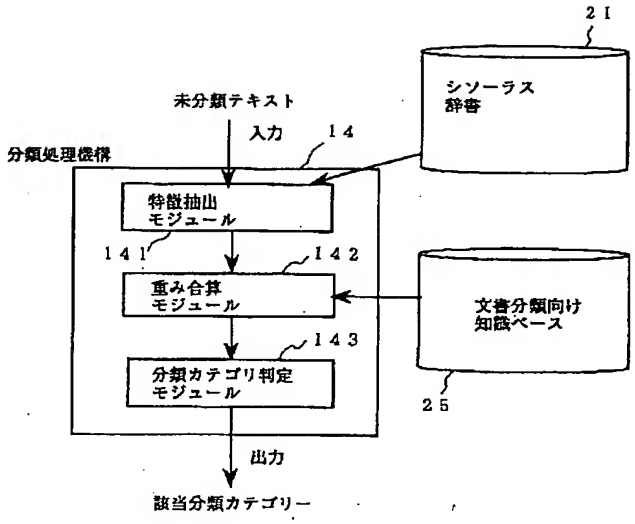
【図７】


特徴ノード
重み
分類カテゴリー
ノード
f1
f2
f3
W1
W2
Wn
+/-
Σ fiwi > θ → +
Σ fiwi < θ → -

【図８】


24
多義解消後
学習データファイル
文書分類向け
知識ベース生成機構
13
131
重み初期設定
モジュール
132
重み合算
モジュール
133
正解判定
モジュール
134
重み更新
モジュール
25
文書分類向け
知識ベース

【図９】


21
シソーラス
辞書
未分類テキスト
入力
14
分類処理機構
特徴抽出
モジュール
141
142
重み合算
モジュール
143
分類カテゴリ判定
モジュール
文書分類向け
知識ベース
25
出力
該当分類カテゴリー